SmartSeq2 quality control report

The data in this document is generated for plate SOGTseqIndex1,SOGTseqIndex2,SOGTseqIndex3.

Additional information regarding quality control can be found in the same folder as this report:

- QC metrics and outlier information per sample
- alignment percentages per sample
- plots used in this document

The kallisto quantifications for each of the samples and merged expression matrices (estimated counts by default) are in the neighbouring directory.

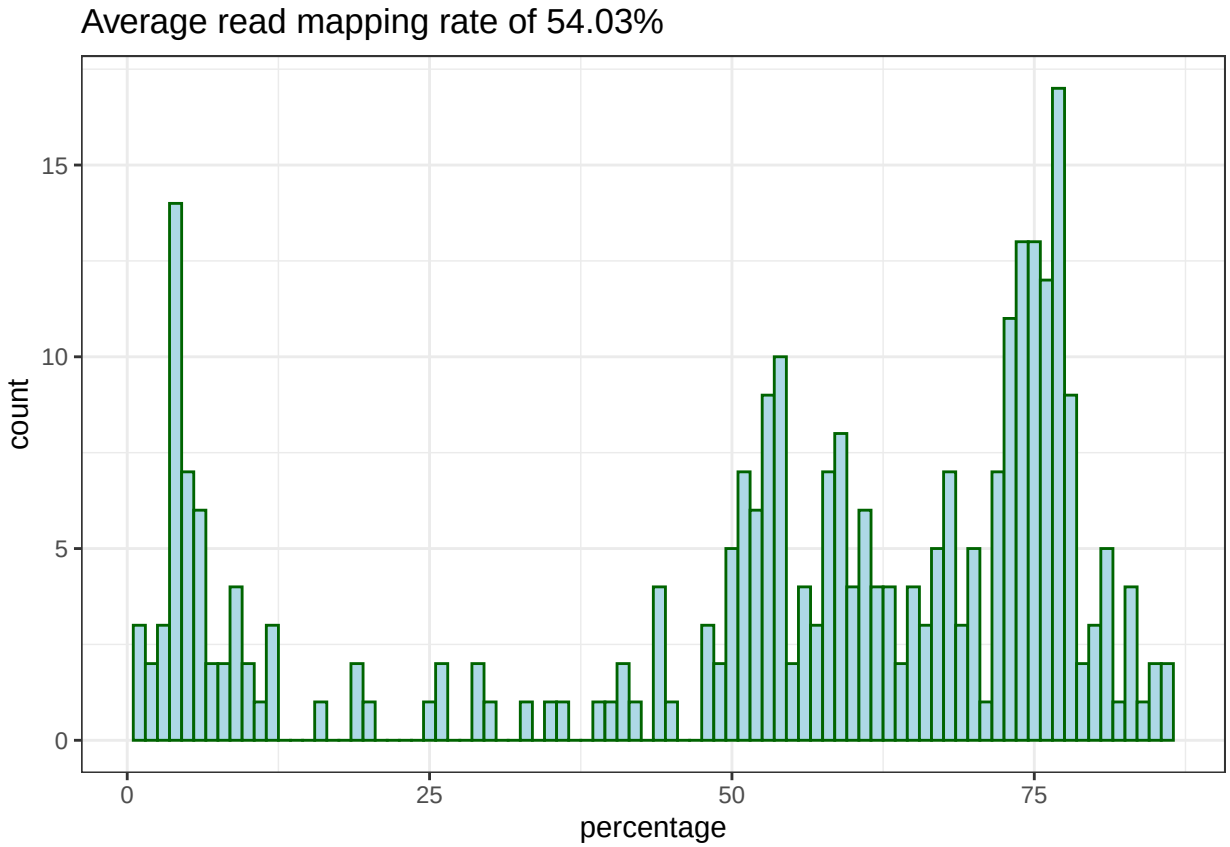
Tools used in the quality control pipeline are:

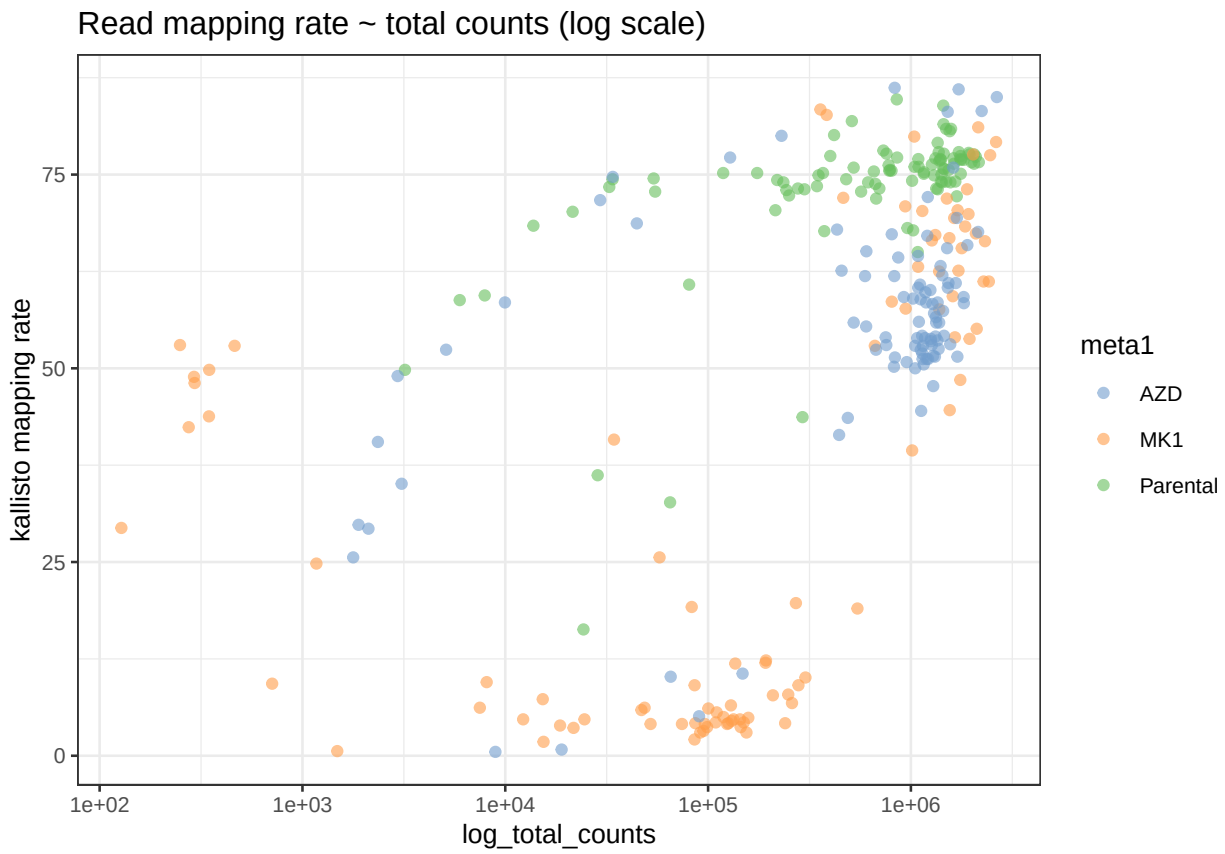
- kallisto 0.44.0
- R 4.1.2
- R markdown
- R packages: scater 1.22.0, ggplot2, dplyr, knitr, rjson
- nextflow 22.04.0
- singularity 3.8.7

Read mapping

Reads were mapped with the kallisto pseudo-aligner to the reference genome.

Percentage of reads that map is shown for all the samples. Samples labelled as control are separated for comparison.

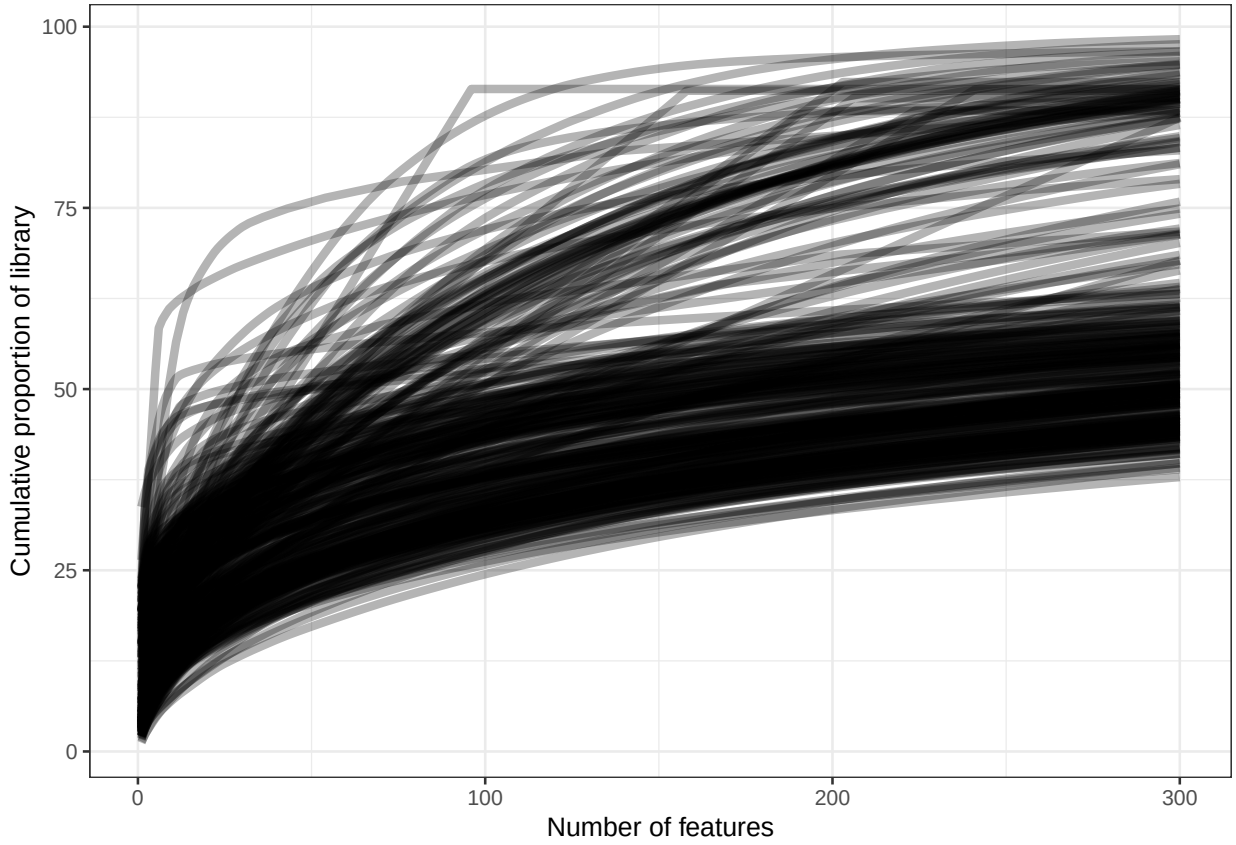

Average read mapping rate of 54.03%
count
15
10
5
0
0
25
50
75
percentage


Read mapping rate ~ total counts (log scale)
kallisto mapping rate
75
50
25
0
1e+02
1e+03
1e+04
1e+05
1e+06
log_total_counts
meta1
AZD
MK1
Parental

Cumulative distribution of expression

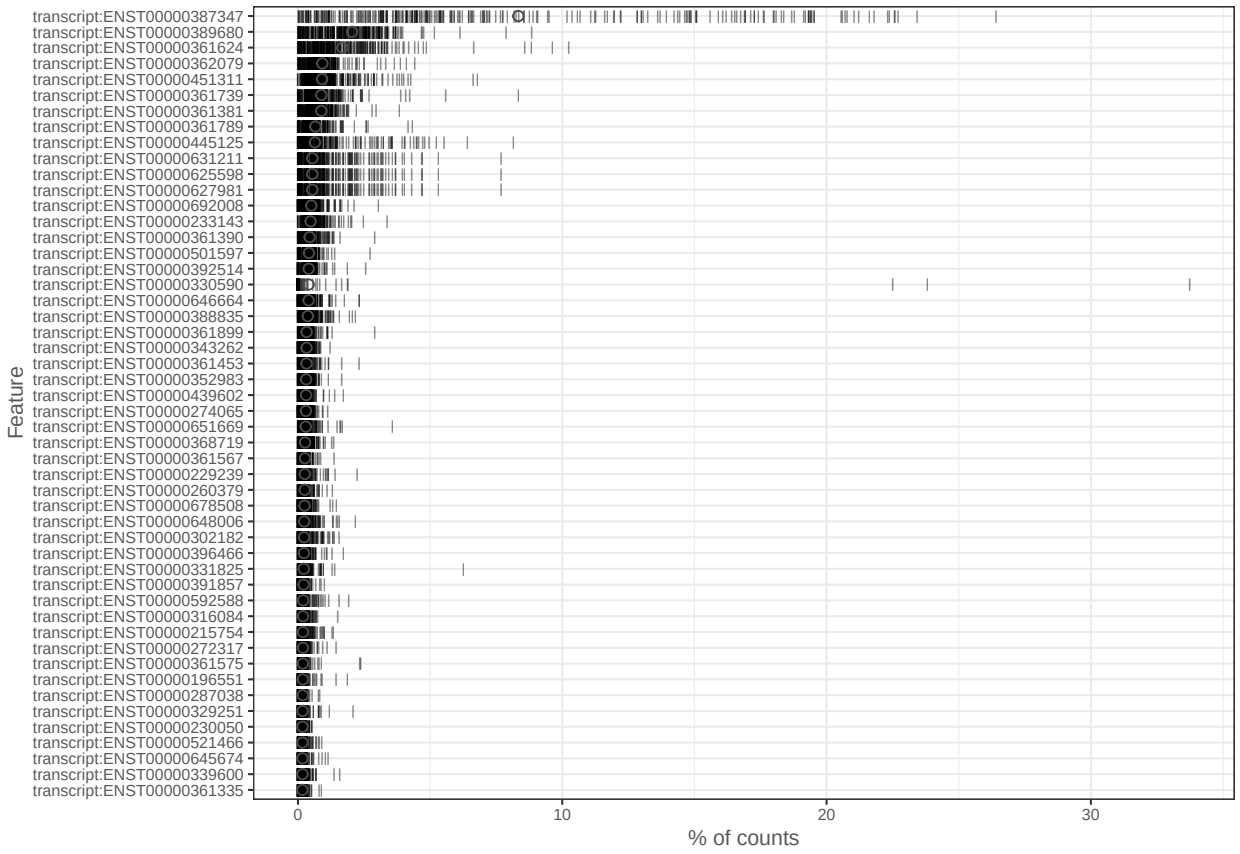
The following plot shows the proportion of the library (y-axis) covered by the top 300 most expressed transcripts (x-axis) for the given sample.

Distributions which rise high quickly are samples which are dominated by low number of transcripts while the ones having a less steep rise have counts more evenly distributed. Wells with less diverse count distribution are more similar to empty wells and might indicate lower quality or damaged cells.



Most highly expressed transcripts

Shown are top 50 features by the proportion of counts they take in all samples. Feature names (as Ensembl transcripts or annotated control features) denote samples on the y-axis with the percentage of counts they capture on the x-axis. A circle shows the mean proportion across all samples with proportion for each of samples shown on the same line.

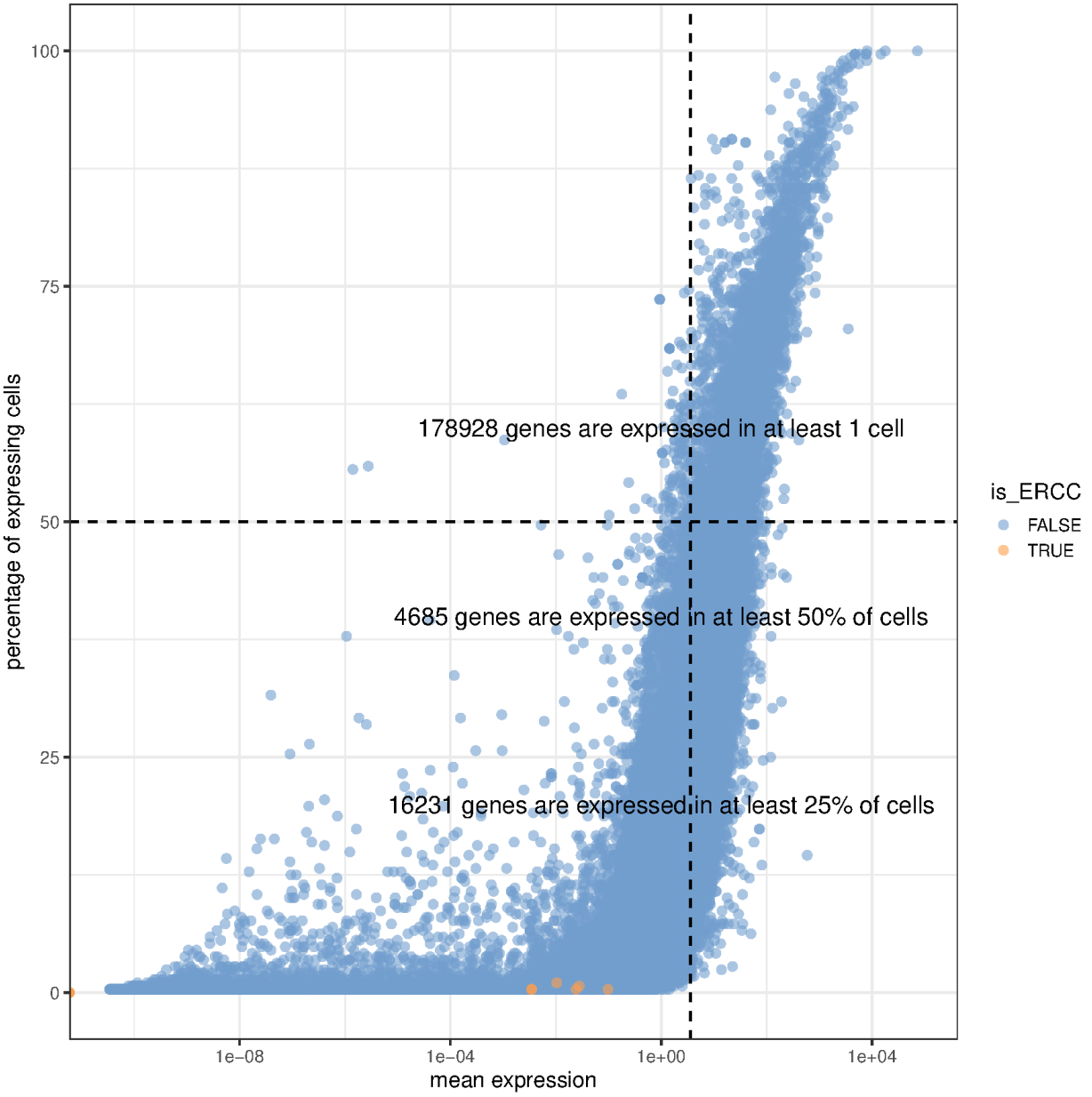


Expression frequency - mean distribution of transcripts

For all of the transcripts (features), their frequency of expression (percentage of samples expressing the specific transcript) is shown against their mean value of expression

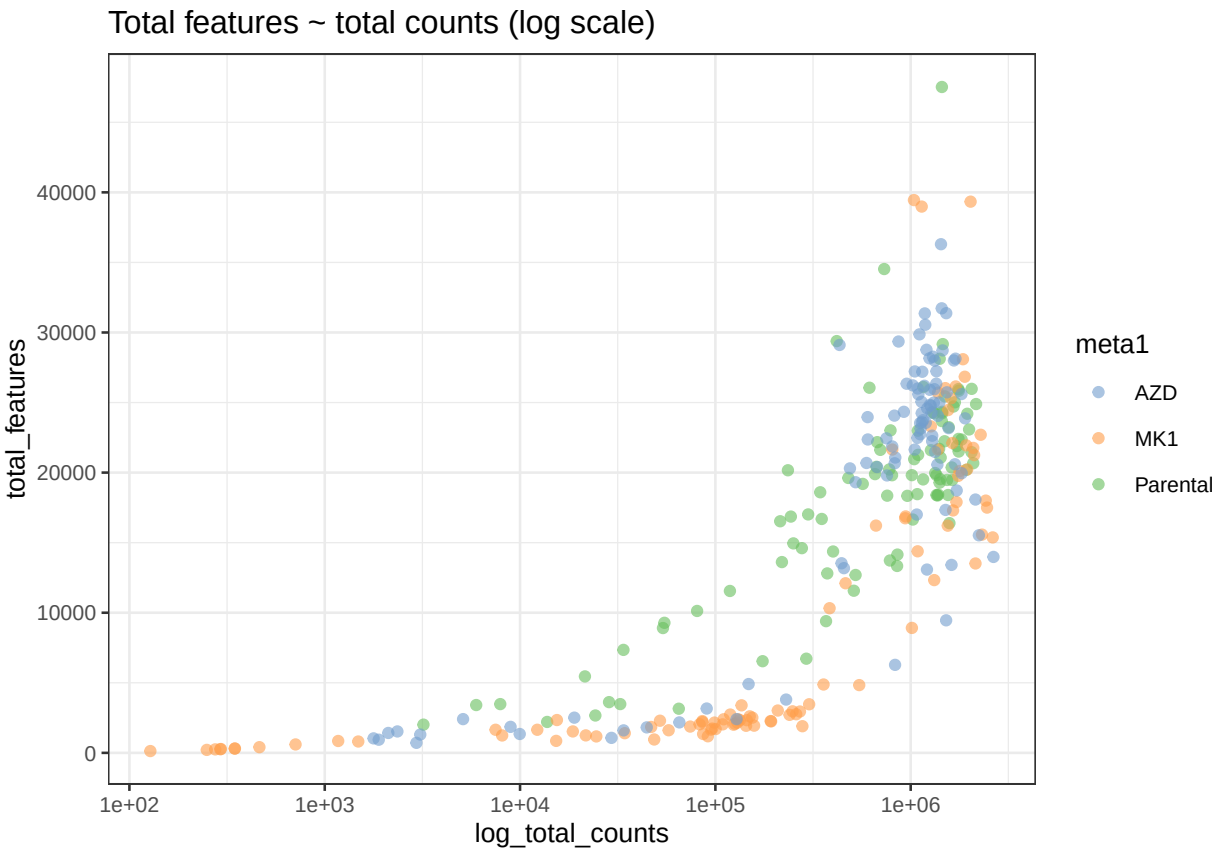
The relationship between the two variables is typically sigmoidal looking.

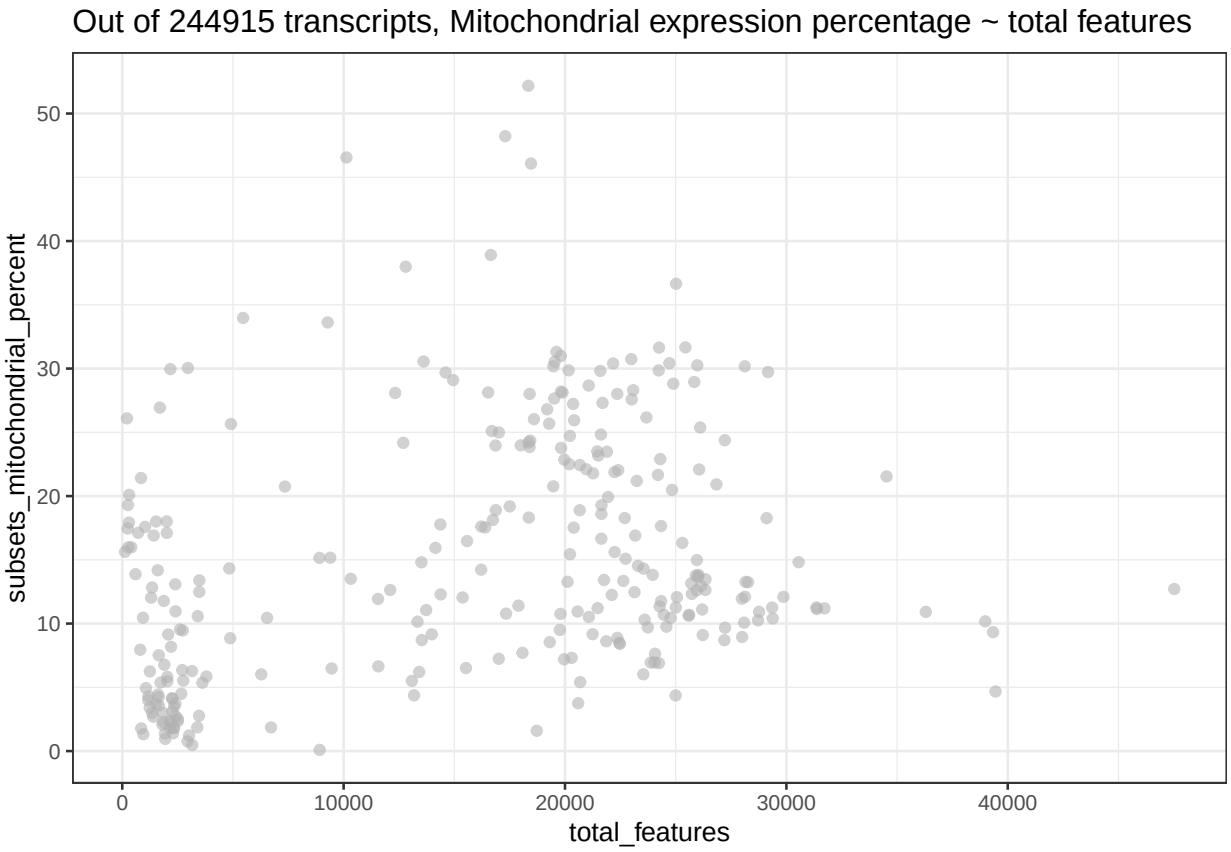
The vertical dashed line is the median of expression levels across the samples. The horizontal dashed line is at 50% of expression presence - dots above are transcripts which are expressed in more than 50% of the samples.

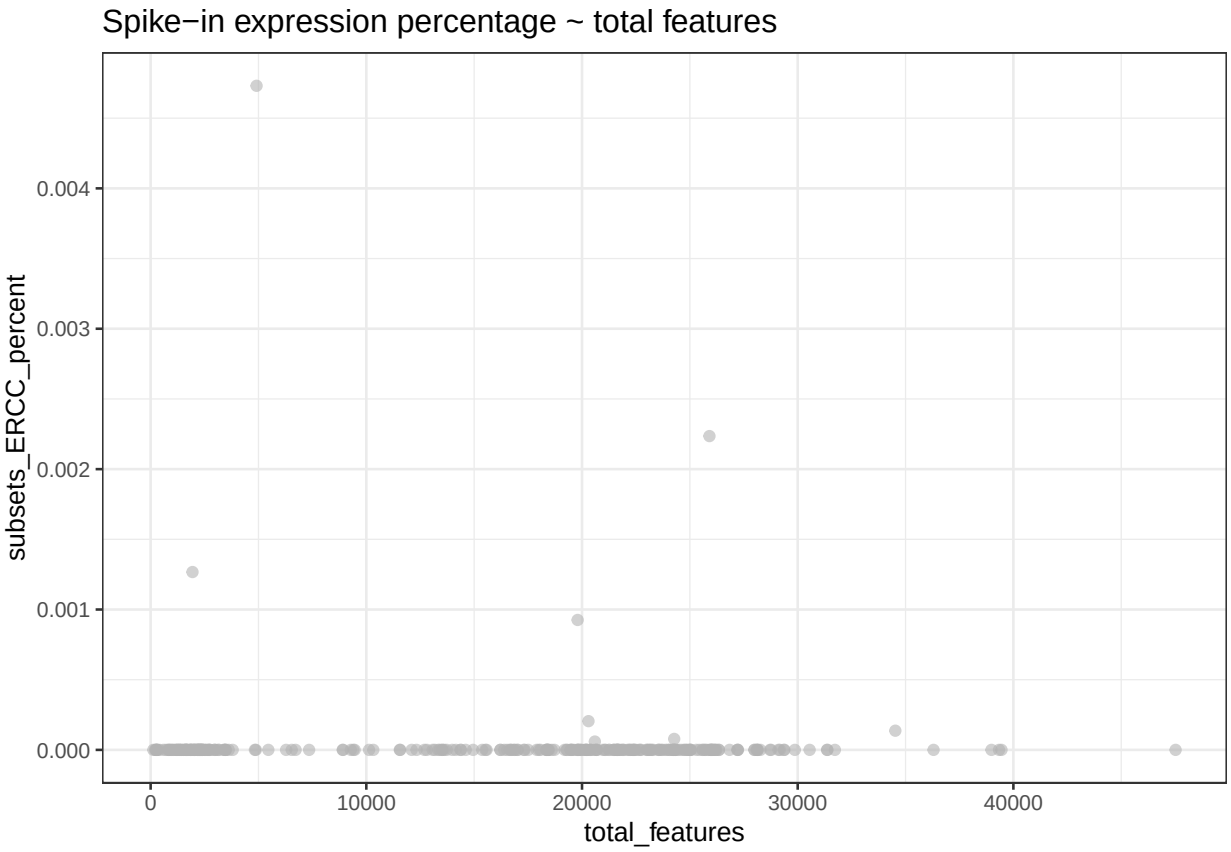


Scatterplots

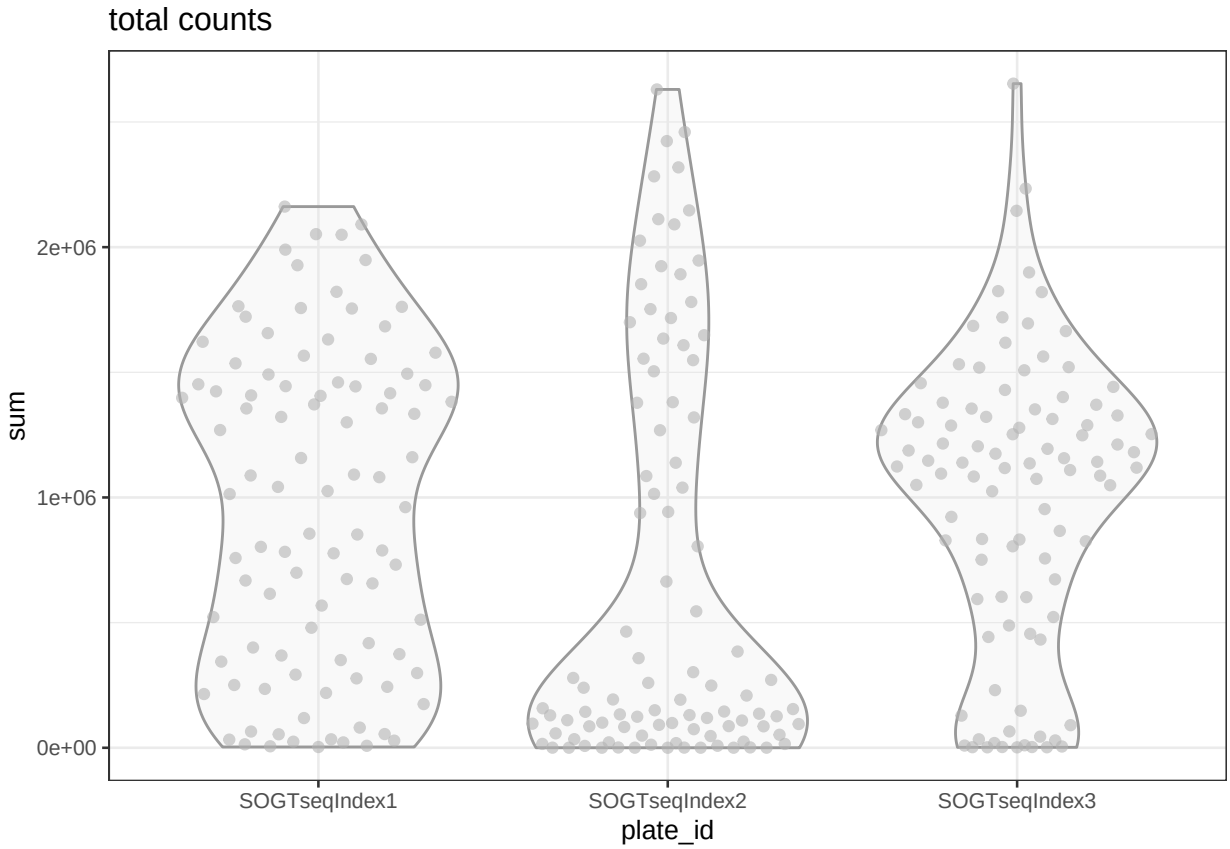
Following scatterplots show, for all of the samples and their total counts, their transcript number, proportion of mitochondrial transcript and proportion of ERCC transcript.

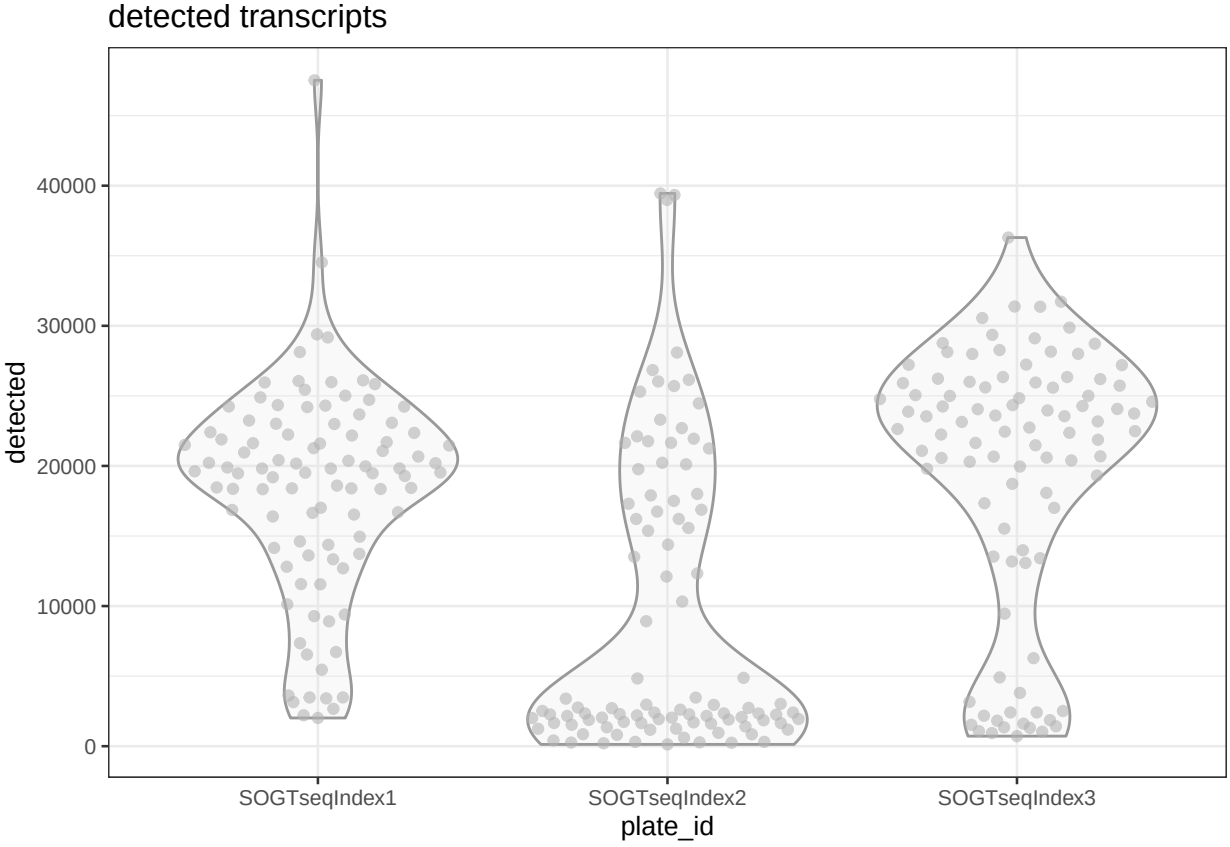


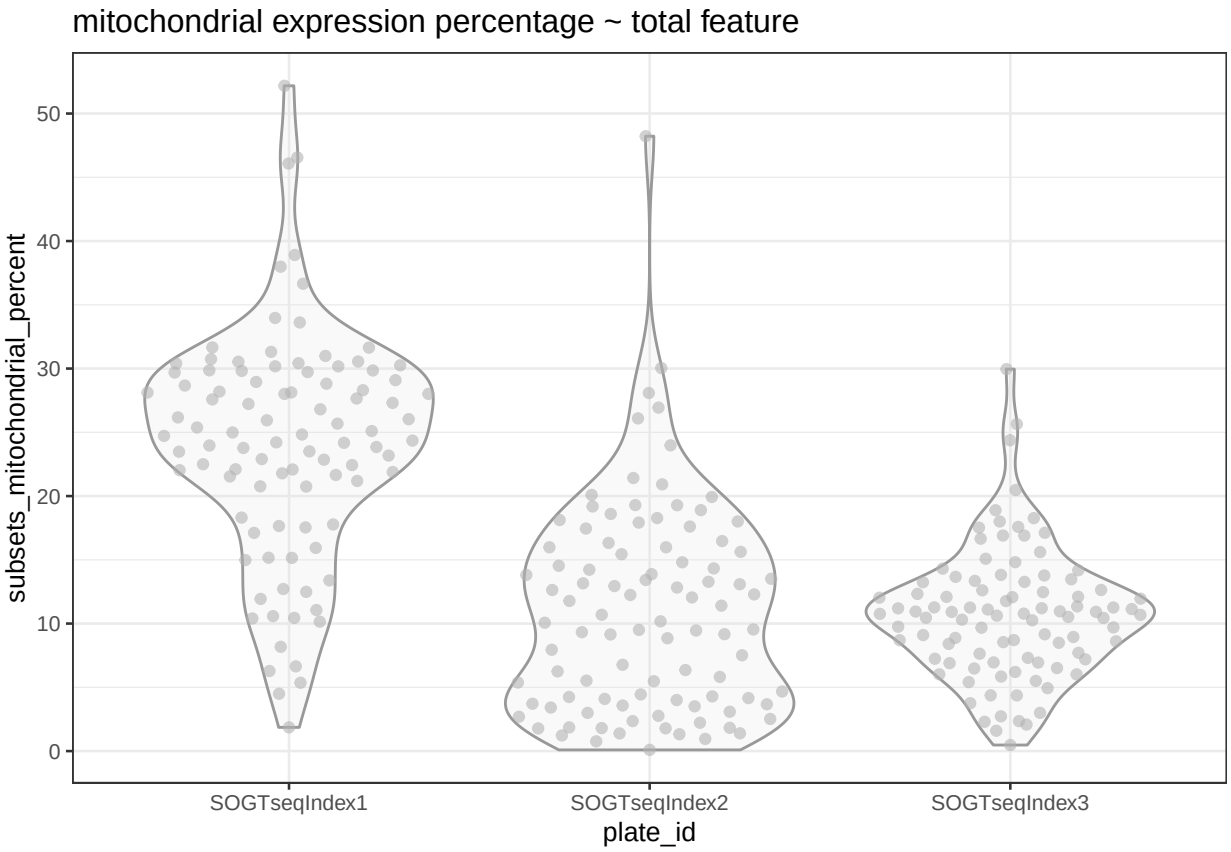

Out of 244915 transcripts, Mitochondrial expression percentage ~ total features
subsets_mitochondrial_percent
50
40
30
20
10
0
0
10000
20000
30000
40000
total_features

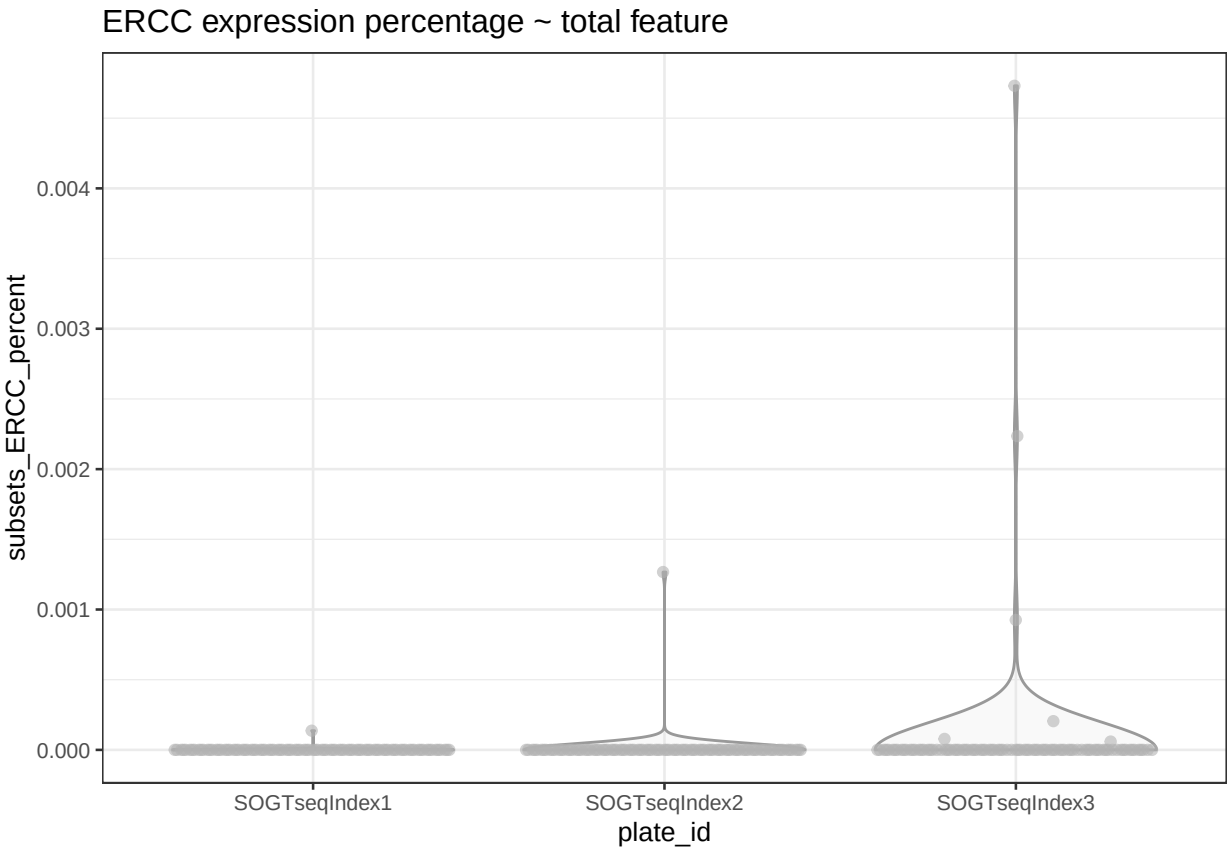

Spike−in expression percentage ~ total features
subsets_ERCC_percent
0.004
0.003
0.002
0.001
0.000
0
10000
20000
30000
40000
total_features

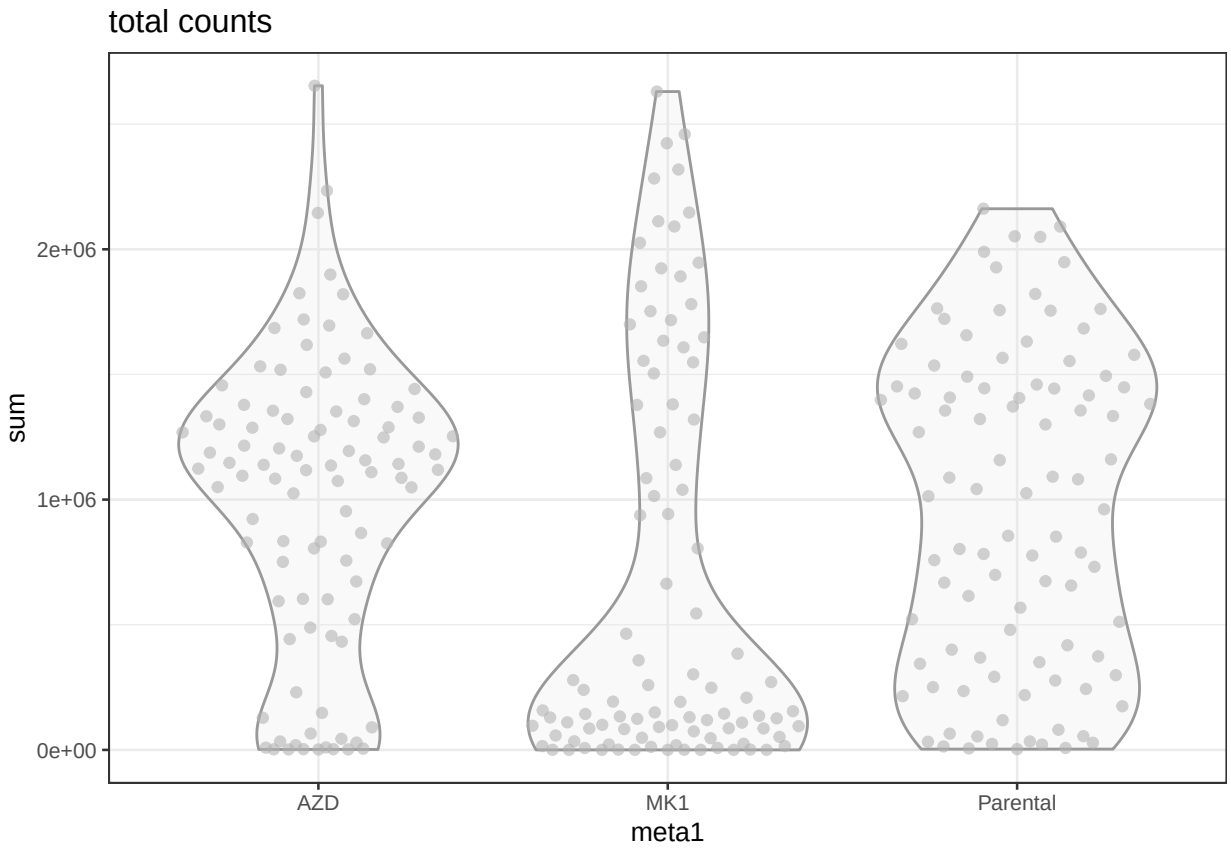
Violin plots

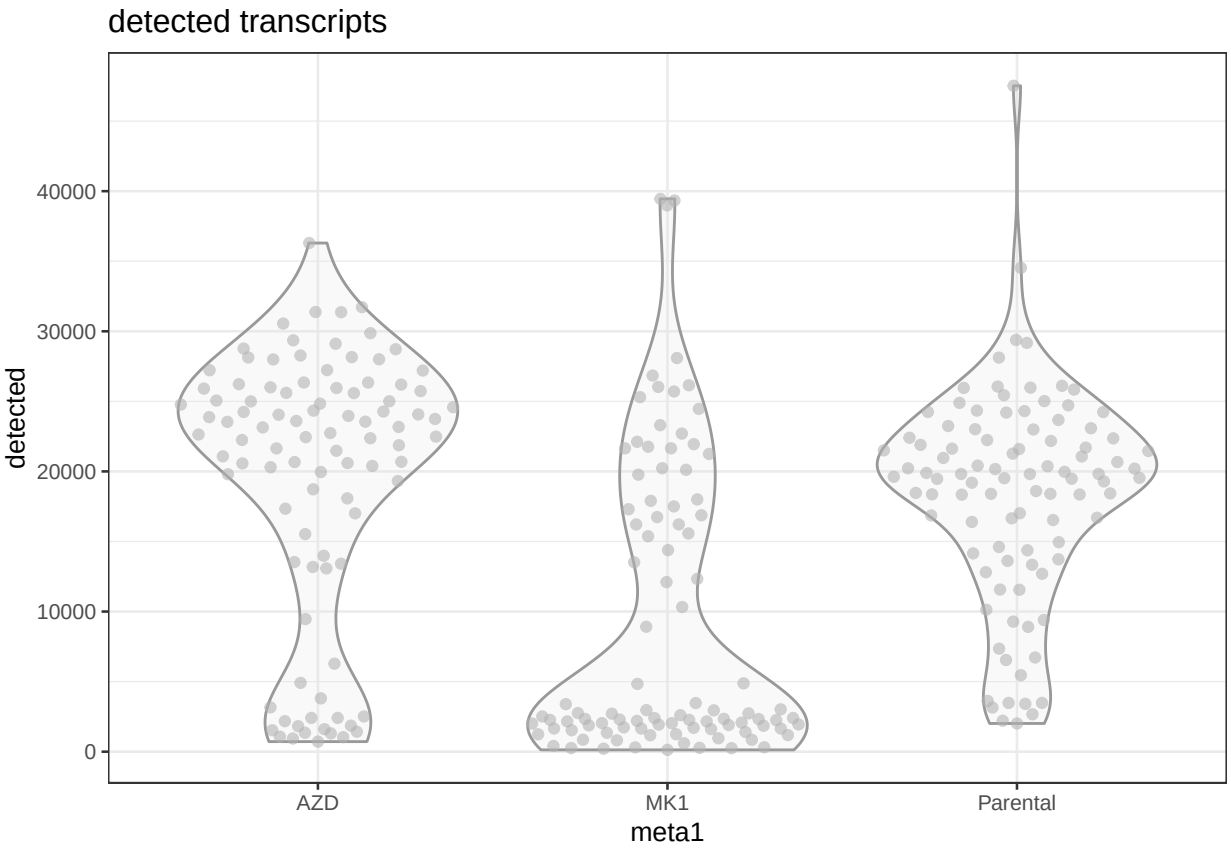


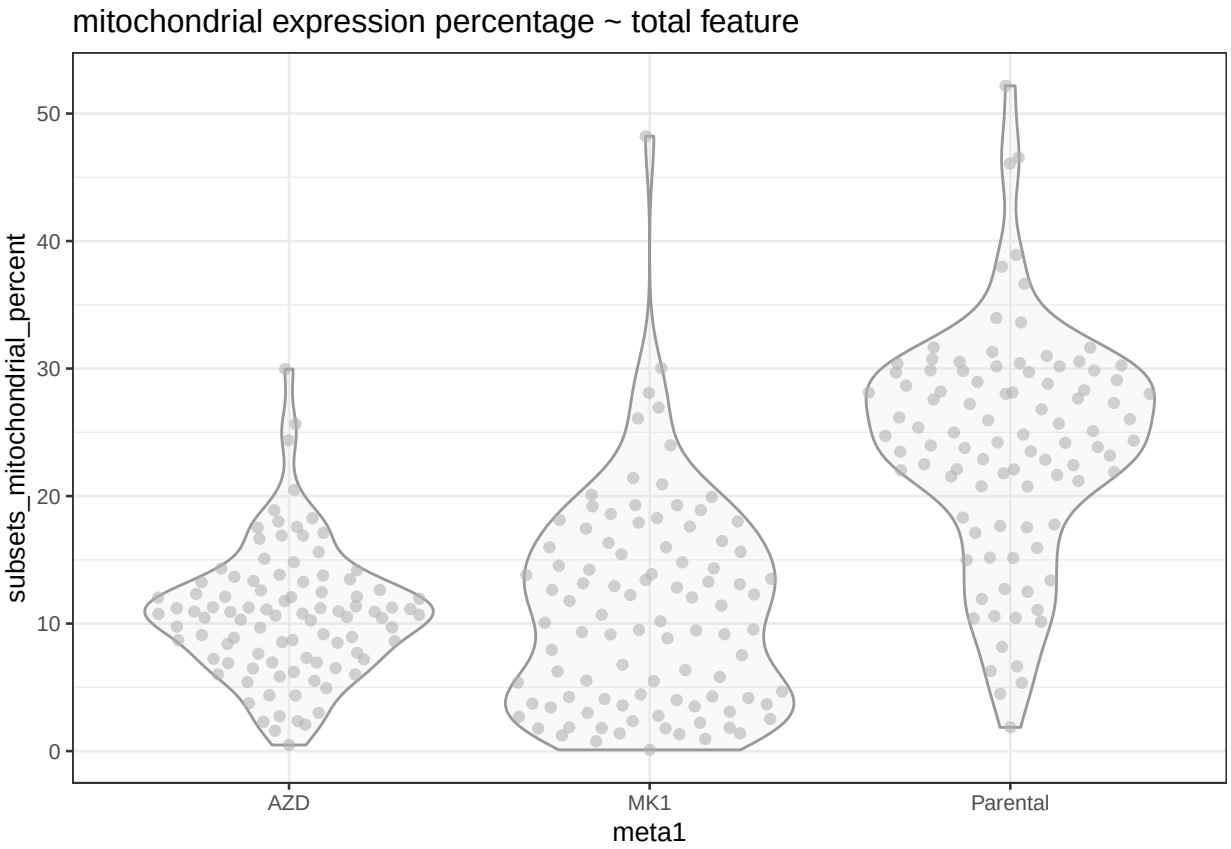


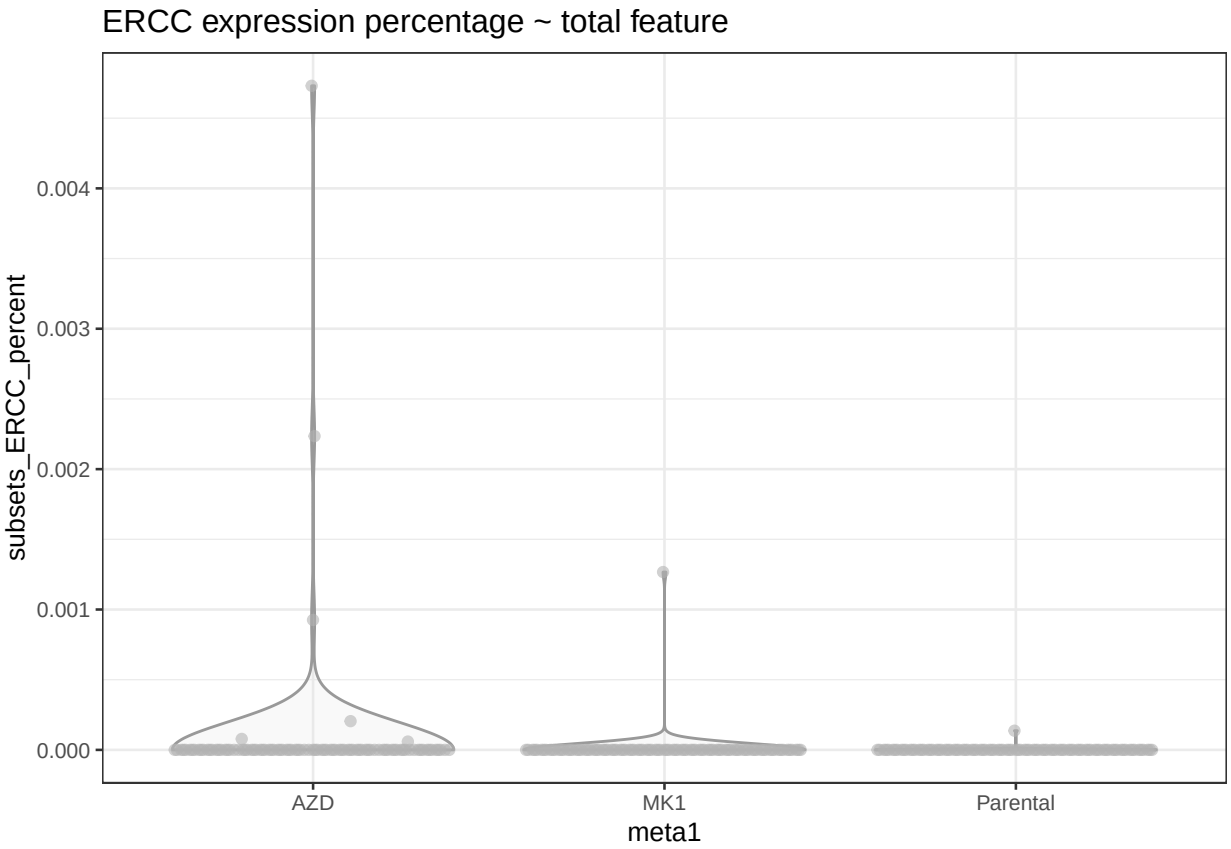

mitochondrial expression percentage ~ total feature
subsets_mitochondrial_percent
50
40
30
20
10
0
SOGTseqIndex1
SOGTseqIndex2
SOGTseqIndex3
plate_id


ERCC expression percentage ~ total feature
subsets_ERCC_percent
0.004
0.003
0.002
0.001
0.000
SOGTseqIndex1
SOGTseqIndex2
SOGTseqIndex3
plate_id


total counts
sum
2e+06
1e+06
0e+00
AZD
MK1
Parental
meta1


detected transcripts
40000
30000
20000
10000
0
detected
AZD
MK1
Parental
meta1


mitochondrial expression percentage ~ total feature
subsets_mitochondrial_percent
50
40
30
20
10
0
AZD
MK1
Parental
meta1


ERCC expression percentage ~ total feature
subsets_ERCC_percent
0.004
0.003
0.002
0.001
0.000
AZD
MK1
Parental
meta1

Dimensionality reductions

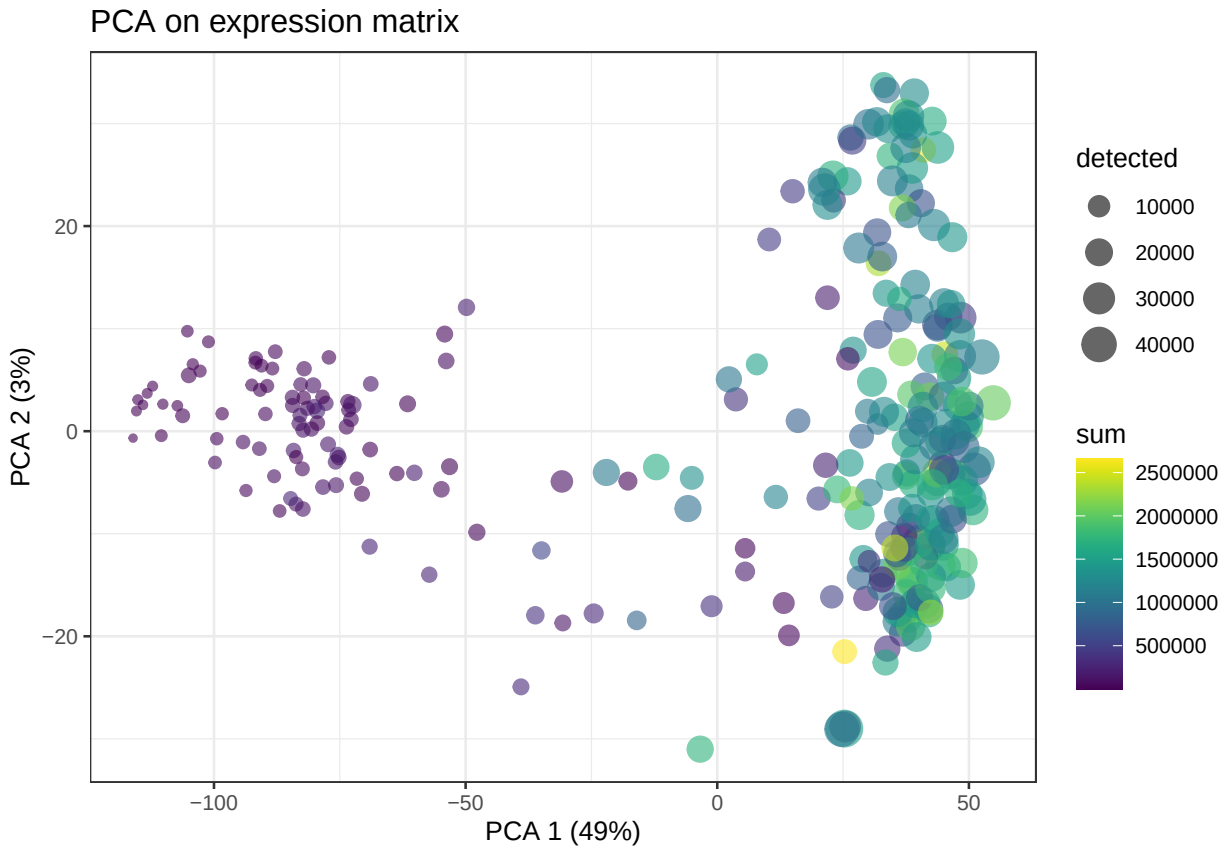
Dimensionality reductions summarise the large feature space with a smaller number of dimensions aiming to capture most of the variance in the data.

Principal component analysis (PCA) and t-SNE are one of the most common methods. PCA makes new dimensions by trying to capture as much as variance as possible in the top ones, while t-SNE tries to place the samples in 2 dimensions in a way which best captures the similarities and differences of the samples.

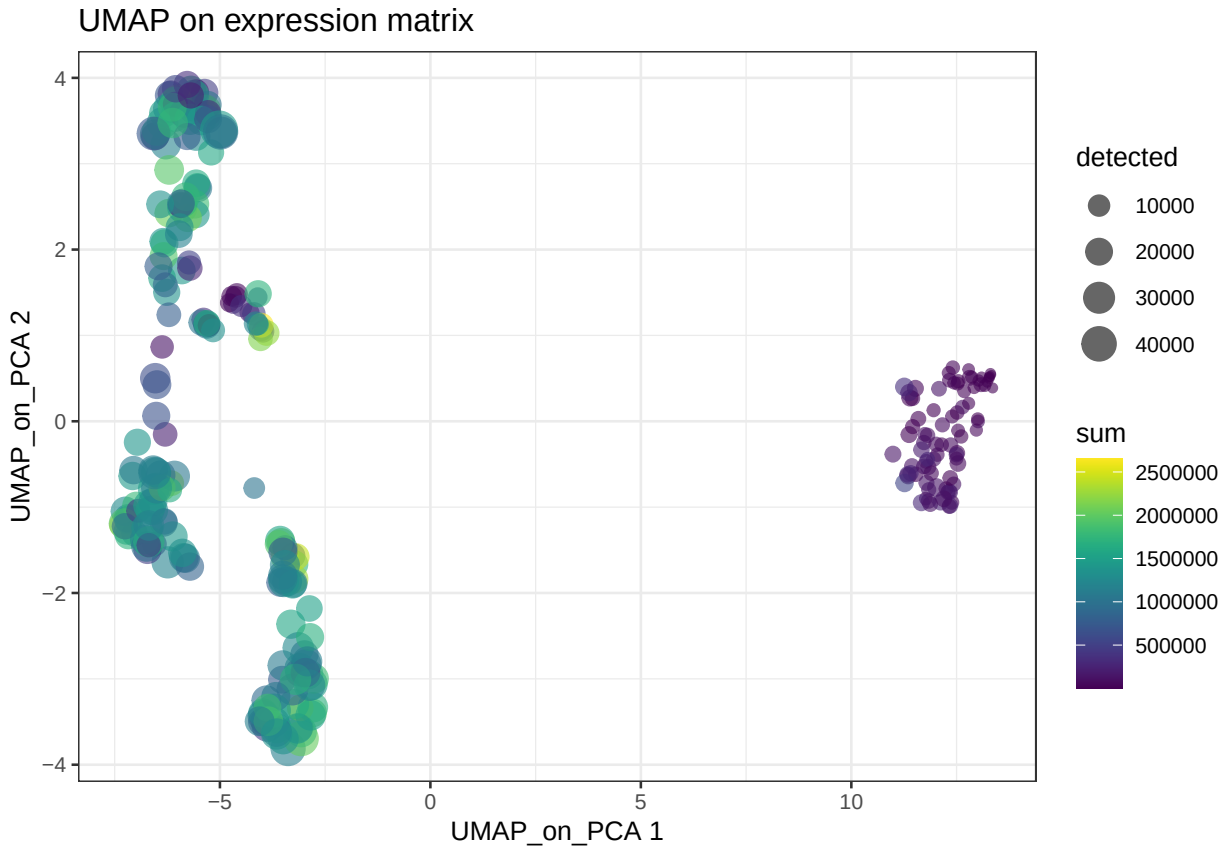
PCA/t-SNE dimensions are derived from the expression profiles of all the transcripts and visualised in the following scatterplots.

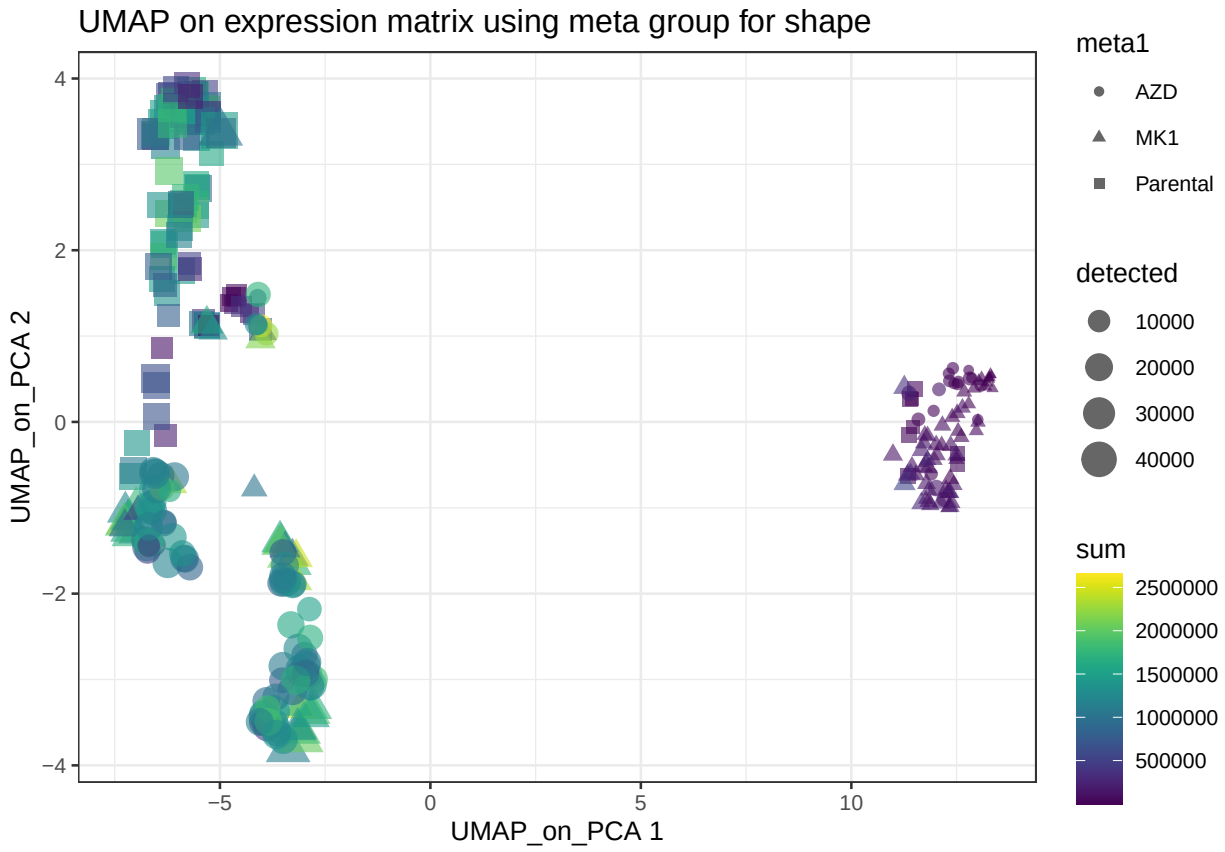
Samples similar in expression profile will be plotted closer together.

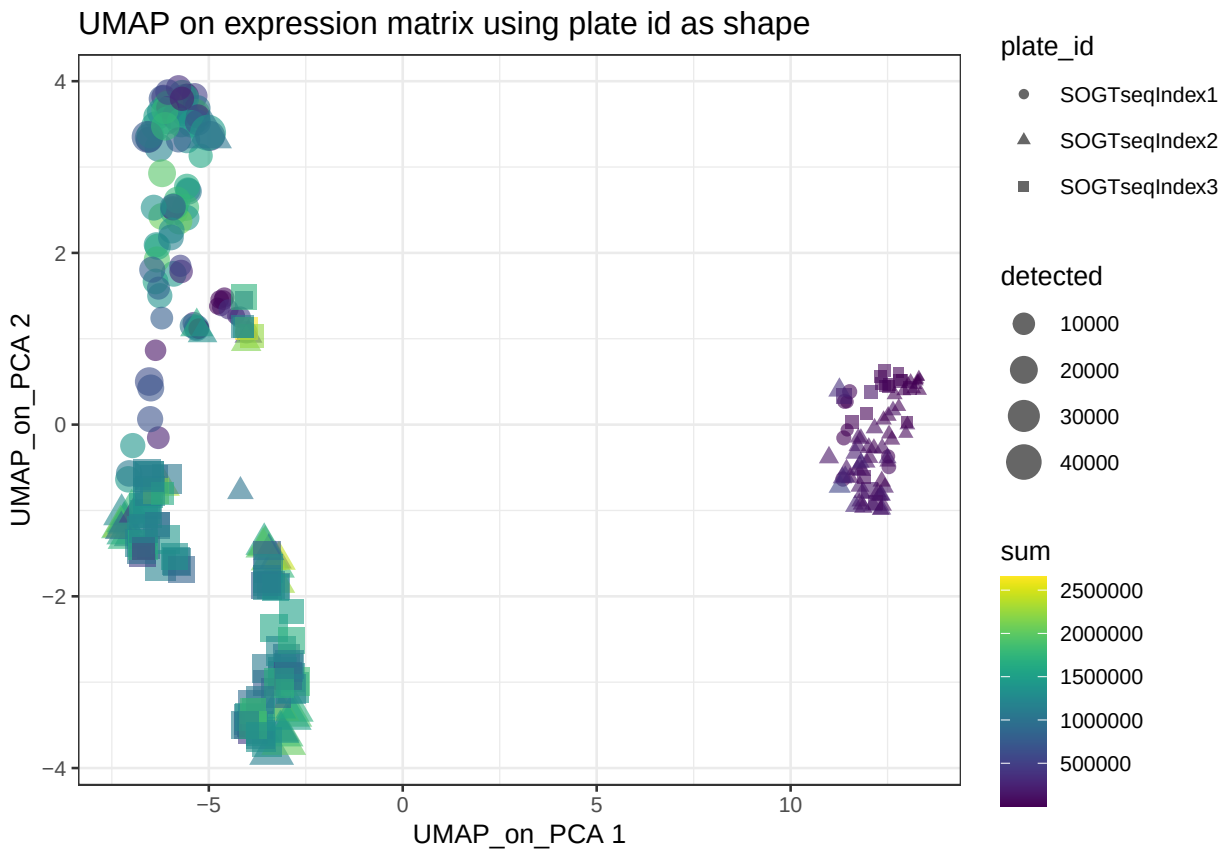
PCA



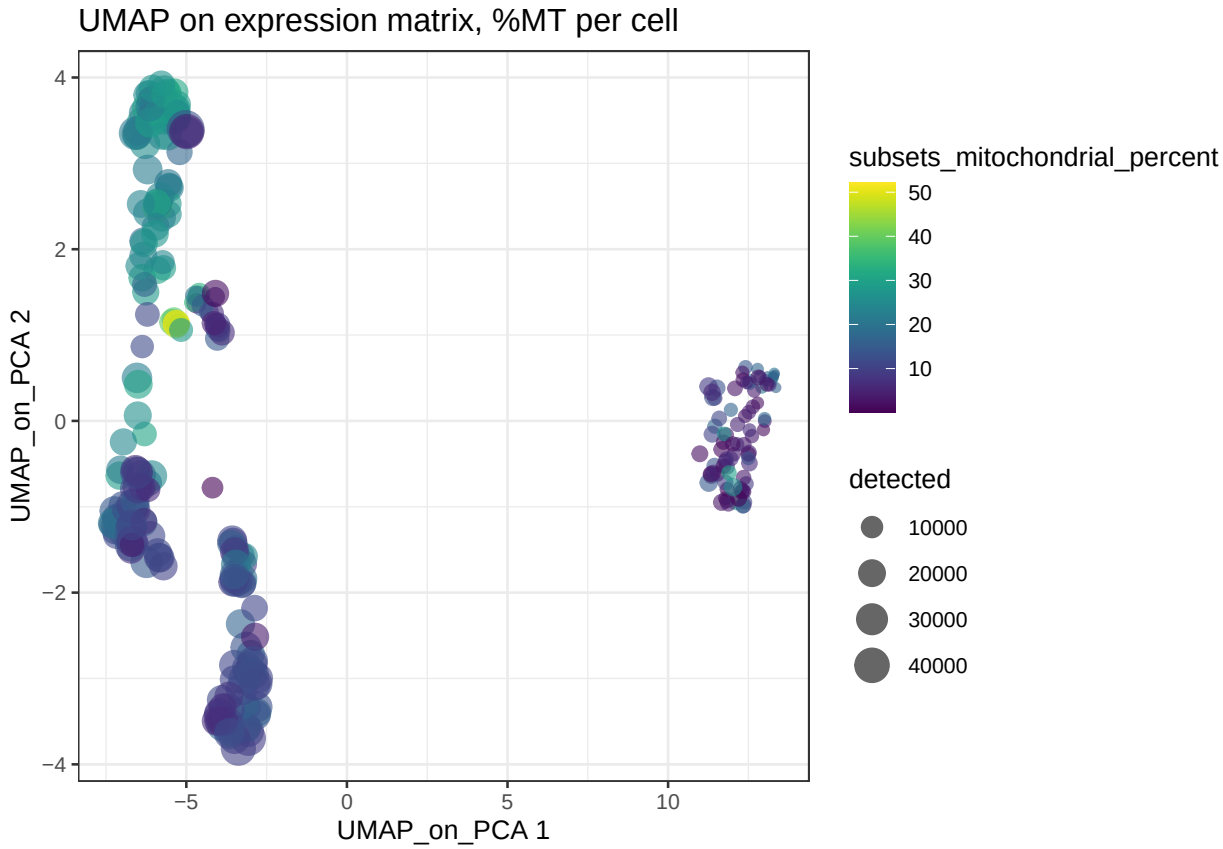
UMAP

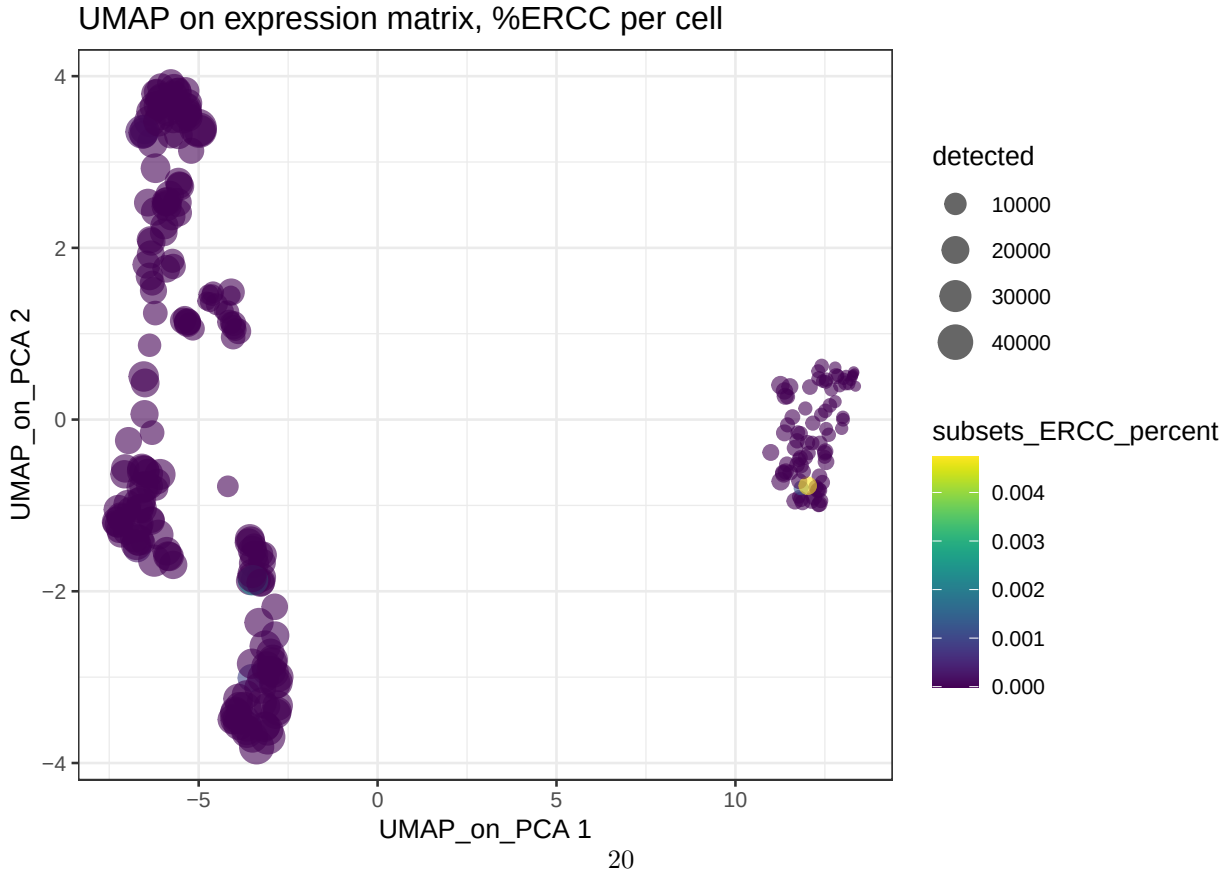

UMAP on expression matrix
UMAP_on_PCA 2
UMAP_on_PCA 1
4
2
0
−2
−4
−5
0
5
10
detected
10000
20000
30000
40000
sum
2500000
2000000
1500000
1000000
500000


UMAP on expression matrix using meta group for shape
UMAP_on_PCA 2
UMAP_on_PCA 1
4
2
0
−2
−4
−5
0
5
10
meta1
AZD
MK1
Parental
detected
10000
20000
30000
40000
sum
2500000
2000000
1500000
1000000
500000


UMAP on expression matrix using plate id as shape
UMAP_on_PCA 2
UMAP_on_PCA 1
4
2
0
−2
−4
−5
0
5
10
plate_id
SOGTseqIndex1
SOGTseqIndex2
SOGTseqIndex3
detected
10000
20000
30000
40000
sum
2500000
2000000
1500000
1000000
500000

UMAP on MT/ERCC





Outliers and key metrics per sample

The accompanying spreadsheet (`Per_sample_key_metrics.tsv`) contains the value for some of the most important metrics, position on the plate, control information and whether or not the sample is considered an outlier in the population when it comes to mapping rate, and percentage of counts which are mitochondrial.

If a sample's average read mapping rate is below 50%, or the mitochondrial percentage is over 10%, the sample is consider as an outlier.

Being an outlier on one metric is unlikely to be of much concern, but if the same sample is an outlier in both metrics, it could be a sign of an unviable sample.

The following samples have been marked as **outliers**.

For 2 outlier metrics:

R0649-S0001_A57839_SOGTseqIndex1G1, R0649-S0001_A57839_SOGTseqIndex2A11, R0649-S0001_A57839_SOGTseqIndex2A6, R0649-S0001_A57839_SOGTseqIndex2B12, R0649-S0001_A57839_SOGTseqIndex2B9, R0649-S0001_A57839_SOGTseqIndex2C5, R0649-S0001_A57839_SOGTseqIndex2C6, R0649-S0001_A57839_SOGTseqIndex R0649-S0001_A57839_SOGTseqIndex2D12, R0649-S0001_A57839_SOGTseqIndex2E12, R0649-S0001_A57839_SOGTseqIndex2F12, R0649-S0001_A57839_SOGTseqIndex2F2, R0649-S0001_A57839_SOGTseqIndex2G12, R0649-S0001_A57839_SOGTseqIndex2H12, R0649-S0001_A57839_SOGTseqIndex2H5, R0649-S0001_A57839_SOGTseqIndex2H7, R0649-S0001_A57839_SOGTseqIndex2H8, R0649-S0001_A57839_SOGTseqIndex3B2, R0649-S0001_A57839_SOGTseqIndex3C12, R0649-S0001_A57839_SOGTseqIndex3C4, R0649-S0001_A57839_SOGTseqIndex3E1, R0649-S0001_A57839_SOGTseqIndex3E9, R0649-S0001_A57839_SOGTseqIndex3F1, R0649-S0001_A57839_SOGTseqIndex3F5, R0649-S0001_A57839_SOGTseqIndex3G12, R0649-S0001_A57839_SOGTseqIndex3G8, R0649-S0001_A57839_SOGTseqIndex3H8

Exact information for the metrics used in the generation of this document can be found in the accompanying table - `Per_sample_key_metrics.tsv`

By default, it contains the following information:

- plate_id
- fastqname
- plate_position
- outlier_hits - the number of times that sample has been labeled as an outlier
- outlier_pct_MT - TRUE if a sample meets the outlier metric on mitochondrial percentage
- outlier_mapping reate - TRUE if a sample meets the outlier metric on average read mapping rate
- sum - total counts
- detected - total expressed features (transcripts)
- subsets_mitochondrial_percent - Percentage of counts belonging to mitochondrial transcripts
- subsets_ERCC_percent - Percentage of counts belonging to spike-ins
- map_rate - average read mapping rate
- is_cell_control